

# Full solution for the storage of correlated memories in an autoassociative memory

Emilio Kropff\*

SISSA - International School of Advanced Studies  
via Beirut 4  
34014, Trieste  
Italy

February 1, 2008

## Abstract

We complement our previous work [Kropff and Treves, 2007] with the full (non diluted) solution describing the stable states of an attractor network that stores correlated patterns of activity. The new solution provides a good fit of simulations of a network storing the feature norms of McRae and colleagues [McRae et al., 2005], experimentally obtained combinations of features representing concepts in semantic memory. We discuss three ways to improve the storage capacity of the network: adding uninformative neurons, removing informative neurons and introducing popularity-modulated hebbian learning. We show that if the strength of synapses is modulated by an exponential decay of the *popularity* of the pre-synaptic neuron, any distribution of patterns can be stored and retrieved with approximately an optimal storage capacity - i.e.  $C_{min} \propto I_{fp}$ , the minimum number of connections per neuron needed to sustain the retrieval of a pattern is proportional to the information content of the pattern multiplied by the number of patterns stored in the network.

## 1 Introduction

Autoassociative memory networks can store patterns of neural activity by modifying the synaptic weights that inter-connect neurons [Hopfield, 1982, Amit, 1989], following the Hebbian rule [Hebb, 1949]. Once a pattern of activity is stored, it becomes an attractor of the dynamics of the system. Direct evidence showing attractor behavior in the hippocampus of *in vivo* animals has been reported [Wills et al., 2005]. These kind of memory systems have been proposed to be present at all levels along the cortex of higher order brains, where hebbian plasticity plays a major role.

Most models of autoassociative memory studied in literature store patterns that are obtained from some random distribution. Some exceptions appeared during the 80's when interest grew around the storage of patterns derived from hierarchical trees [Parga and Virasoro, 1986, Gutfreund, 1988]. Of particular interest, Virasoro [Virasoro, 1988] relates the behavior of networks of general architecture with *prosopagnosia*, an impairment that impedes a patient to individuate certain stimuli without affecting its capacity to categorize them. Interestingly, the results from this model indicate that prosopagnosia is not present in Hebbian-plasticity derived networks. Some other developments have used perceptron-like or other arbitrary local rules for storing generally correlated patterns [Gardner et al., 1989, Diederich and Oppen, 1987] or patterns with spatial correlation [Monasson, 1992]. More recently, Tsodyks and collaborators [Blumenfeld et al., 2006] have studied a Hopfield memory in which a sequence of morphs between two uncorrelated patterns are stored. In this work, the use of a saliency function favouring unexpected over expected patterns during learning results in the formation of a continuous one-dimensional attractor that spans the space between the two original memories. The fusion of basins of attraction can be an interesting phenomenon that we are not going to

---

\*kropff@sisssa.it - <http://people.sissa.it/~kropff>

treat in this work, since we assume that the elements stored in a memory such as the semantic one are differentiable by construction.

Feature norms are a way to get an insight on how semantic information is organized in the human brain [Vinson and Vigliocco, 2002, Garrard et al., 2001, McRae et al., 2005]. The information is collected by asking different types of questions about particular concepts to a large population of subjects. Representations of the concepts are obtained in terms of the features that appear more often in the subjects' descriptions. In this work we analyze the feature norms of McRae and colleagues [McRae et al., 2005] for two reasons: they are public and the size of the dataset allows a statistical approach (it includes 541 concepts described in terms of 2526 features). The norms were downloaded from the *Psychonomic Society Archive of Norms, Stimuli, and Data* web site ([www.psychonomic.org/archive](http://www.psychonomic.org/archive)) with consent of the authors.

In section 2 we define a simple binary associative network, showing how it can be modified in order to store correlated representations. In section 3 we solve the equilibrium equations for the stable attractor states of the system using a self-consistent signal to noise approach. Finally, in section 4 we study the storage of the feature norms of McRae and colleagues representing semantic memory elements.

## 2 The model

We assume a network with  $N$  neurons and  $C \leq N$  synaptic connections per neuron. If the network stores  $p$  patterns, the parameter  $\alpha = p/C$  is a measure of the memory load normalized by the size of the network. In classical models, the equilibrium properties of large enough networks depends on  $p$ ,  $C$  and  $N$  only through  $\alpha$ , which allows the definition of the thermodynamic limit ( $p \rightarrow \infty$ ,  $C \rightarrow \infty$ ,  $N \rightarrow \infty$ ,  $\alpha$  constant).

The activity of neuron  $i$  is described by the variable  $\sigma_i$ , with  $i = 1 \dots N$ . Each of the  $p$  patterns is a particular state of activation of the network. The activity of neuron  $i$  in pattern  $\mu$  is described by  $\xi_i^\mu$ , with  $\mu = 1 \dots p$ . The perfect retrieval of pattern  $\mu$  is thus characterized by  $\sigma_i = \xi_i^\mu$  for all  $i$ . We will assume binary patterns, where  $\xi_i^\mu = 0$  if the neuron is silent and  $\xi_i^\mu = 1$  if the neuron fires. Consistently, the activity states of neurons will be limited by  $0 \leq \sigma_i \leq 1$ . We will further assume a fraction  $a$  of the neurons being activated in each pattern. This quantity receives the name of *sparseness*.

Each neuron receives  $C$  synaptic inputs. To describe the architecture of connections we use a random matrix with elements  $C_{ij} = 1$  if a synaptic connection between post-synaptic neuron  $i$  and pre-synaptic neuron  $j$  exists and  $C_{ij} = 0$  otherwise, with  $C_{ii} = 0$  for all  $i$ . In addition to this, synapses have associated weights  $J_{ij}$ .

The influence of the network activity on a given neuron  $i$  is represented by the field

$$h_i = \sum_{j=1}^N C_{ij} J_{ij} \sigma_j \quad (1)$$

which enters a sigmoidal activation function in order to update the activity of the neuron

$$\sigma_i = \{1 + \exp \beta (U - h_i)\}^{-1} \quad (2)$$

where  $\beta$  is inverse to a temperature parameter and  $U$  is a threshold favoring silence among neurons [Buhmann et al., 1989, Tsodyks and Feigl'Man, 1988].

The learning rule that defines the weights  $J_{ij}$  must reflect the Hebbian principle: every pattern in which both neurons  $i$  and  $j$  are active will contribute positively to  $J_{ij}$ . In addition to this, the rule must include, in order to be optimal, some prior information about pattern statistics. In a one-shot learning paradigm, the optimal rule uses the sparseness  $a$  as a learning threshold,

$$J_{ij} = \frac{1}{Ca} \sum_{\mu=1}^p (\xi_i^\mu - a) (\xi_j^\mu - a). \quad (3)$$

However, as we have shown in previous work [Kropff and Treves, 2007], in order to store correlated patterns this rule must be modified using  $a_j$ , or the *popularity* of the pre-synaptic neuron, as a learning threshold,

$$J_{ij} = \frac{1}{Ca} \sum_{\mu=1}^p \xi_i^\mu (\xi_j^\mu - a_j), \quad (4)$$

with

$$a_i \equiv \frac{1}{p} \sum_{\mu=1}^p \xi_i^\mu. \quad (5)$$

This requirement comes from splitting the field into a signal and a noise part,

$$h_i = \frac{1}{Ca} \xi_i^1 \sum_{j=1}^N C_{ij} (\xi_j^1 - a_j) \sigma_j + \frac{1}{Ca} \sum_{\mu=2}^p \xi_i^\mu \sum_{j=1}^N C_{ij} (\xi_j^\mu - a_j) \sigma_j, \quad (6)$$

and, under the hypothesis of gaussian noise, setting the average to zero and minimizing the variance. This last is

$$\begin{aligned} var = & \frac{1}{C^2 a^2} \sum_{\mu=1}^p \xi_i^\mu \sum_{j=1}^N C_{ij} \sigma_j^2 (\xi_j^\mu - a_j)^2 + \\ & + \frac{1}{C^2 a^2} \sum_{\mu \neq \nu=1}^p \xi_i^\mu \xi_i^\nu \sum_{j=1}^N C_{ij} \sigma_j^2 (\xi_j^\mu - a_j) (\xi_j^\nu - a_j) + \\ & + \frac{1}{C^2 a^2} \sum_{\mu=1}^p \xi_i^\mu \sum_{j \neq k=1}^N C_{ij} C_{ik} \sigma_j \sigma_k (\xi_j^\mu - a_j) (\xi_k^\mu - a_k) + \\ & + \frac{1}{C^2 a^2} \sum_{\mu \neq \nu=1}^p \xi_i^\mu \xi_i^\nu \sum_{j \neq k=1}^N C_{ij} C_{ik} \sigma_j \sigma_k (\xi_j^\mu - a_j) (\xi_k^\nu - a_k). \end{aligned} \quad (7)$$

If statistical independence is granted between any two neurons, only the first term in Eq. 7 survives when averaging over  $\{\xi\}$ .

In Figure 1 we show that the rule in Eq. 3 can effectively store uncorrelated patterns taken from the distribution

$$P(\xi_i^\mu) = a \delta(\xi_i^\mu - 1) + (1 - a) \delta(\xi_i^\mu). \quad (8)$$

but cannot handle less trivial distributions of patterns, suffering a storage collapse. The storage capacity can be brought back to normal by using the learning rule in Eq. 4, which is also suitable for storing uncorrelated patterns.

Having defined the optimal model for the storage of correlated memories, we analyze in the following sections the storage properties and its consequences through mean field equations.

### 3 Self consistent analysis for the stability of retrieval

We now proceed to derive the equations for the stability of retrieval, similarly to what we have done in [Kropff and Treves, 2007] but in a network with an arbitrary level of random connectivity, where the approximation  $C \ll N$  is no longer valid [Shiino and Fukai, 1992, Shiino and Fukai, 1993, Roudi and Treves, 2004]. Furthermore, we introduce patterns with variable mean activation, given by

$$d_\mu \equiv \frac{1}{N} \sum_{j=1}^N \xi_j^\mu \quad (9)$$

for a generic pattern  $\mu$ . As a result of this, the optimal weights are given by

$$J_{ij} = g_j \sum_{\mu=1}^p \frac{c_{ij}}{C d_\mu} \xi_i^\mu (\xi_j^\mu - a_j) \quad (10)$$

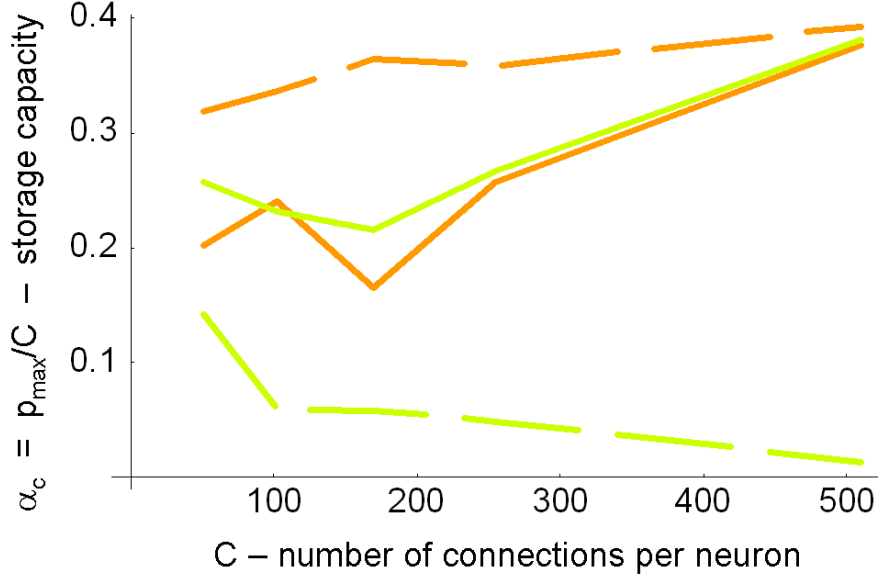


Figure 1: The four combinations of two learning rules and two types of dataset. Green: one shot 'standard' learning rule of Eq. 3. Orange: modified rule of Eq. 4. Solid: trivial distribution of randomly correlated patterns obtained from Eq. 8. Dashed: non-trivially correlated patterns obtained using a hierarchical algorithm. In three cases, the storage capacity (the maximum number of retrievable patterns normalized by  $C$ ) with  $C$  (the number of connections per neuron) is finite and converges to a common value as  $C$  increases. Only in the case of one-shot learning of correlated patterns there is a storage collapse.

which ensures that patterns with different overall activity will have not only a similar noise but also a similar signal. In addition, we have introduced a factor  $g_j = g(a_j)$  in the weights that may depend on the popularity of the pre-synaptic neuron. We will consider  $g_j = 1$  for all but the last section of this work.

If the generic pattern 1 is being retrieved, the field in Eq. 1 for neuron  $i$  can be written as a signal and a noise contribution

$$h_i = \xi_i^1 m_i^1 + \sum_{\mu \neq 1} \xi_i^\mu m_i^\mu \quad (11)$$

with

$$m_i^\mu = \frac{1}{Cd_\mu} \sum_{j=1}^N g_j c_{ij} (\xi_j^\mu - a_j) \sigma_j. \quad (12)$$

We hypothesize that in a stable situation the second term in Eq.11, the noise, can be decomposed into two contributions

$$\sum_{\mu \neq 1} \xi_i^\mu m_i^\mu = \gamma_i \sigma_i + \rho_i z_i. \quad (13)$$

The second term in Eq. 13 represents a gaussian noise with standard deviation  $\rho_i$ , and  $z_i$  a random variable taken from a normal distribution of unitary standard deviation. The first term is proportional to the activity of the neuron  $i$  and results from closed synaptic loops that propagate this activity through the network back to the original neuron, as shown in [Roudi and Treves, 2004]. As is typical in the self consistent method, we will proceed to estimate  $m_i^\mu$  from the ansatz in Eq. 13, inserting it into Eq. 11 and validating the result with, again, Eq. 13, checking the consistency of the ansatz.

Since Eq. 13 is a sum of  $p \rightarrow \infty$  microscopic terms, we can take a single term  $\nu$  out and assume that the sum changes only to a negligible extent. In this way, the field becomes

$$h_i \simeq \xi_i^1 m_i^1 + \xi_i^\nu m_i^\nu + \gamma_i \sigma_i + \rho_i z_i. \quad (14)$$

If the network has reached stability, which we assume, updating neuron  $i$  does not affect its state. This can be expressed by inserting the field into Eq. 2,

$$\sigma_i = \{1 + \exp(-\beta(h_i - U))\}^{-1} \equiv G [\xi_i^1 m_i^1 + \xi_i^\nu m_i^\nu + \rho_i z_i]. \quad (15)$$

In the RHS of Eq. 15 the contribution of  $\gamma_i \sigma_i$  to the field has been reabsorbed into the definition of  $G[x]$ . At first order in  $\xi_j^\nu m_j^\nu$ , Eq. 15 corresponding to neuron  $j$  can be written as

$$\sigma_j \simeq G [\xi_j^1 m_j^1 + \rho_j z_j] + G' [\xi_j^1 m_j + \rho_j z_j] \xi_j^\nu m_j^\nu. \quad (16)$$

To simplify the notation we will further use  $G_j \equiv G [\xi_j^1 m_j^1 + \rho_j z_j]$  and  $G'_j \equiv G' [\xi_j^1 m_j + \rho_j z_j]$ . To this order of approximation, Eq. 12 becomes

$$m_i^\mu = \frac{1}{Cd_\mu} \sum_{j=1}^N N g_j c_{ij} (\xi_j^\mu - a_j) \{G_j + G'_j \xi_j^\mu m_j^\mu\}. \quad (17)$$

Other terms of the same order in the Taylor expansion could have been introduced in Eq. 16, corresponding to the derivatives of  $G$  with respect to  $\xi_j^\mu m_j^\mu$  for  $\mu \neq \nu$ . It is possible to show, however, that such terms give a negligible contribution to the field.

If we define

$$\begin{aligned} L_i^\mu &= \frac{1}{Cd_\mu} \sum_{j=1}^N g_j c_{ij} (\xi_j^\mu - a_j) G_j \\ K_{ij}^\mu &= \frac{1}{Cd_\mu} g_j c_{ij} (\xi_j^\mu - a_j) \xi_j^\mu G'_j, \end{aligned} \quad (18)$$

Eq. 17 can be simply expressed as

$$m_i^\mu = L_i^\mu + \sum_{j=1}^N K_{ij}^\mu m_j^\mu. \quad (19)$$

This equation can be applied recurrently to itself renaming indexes,

$$m_i^\mu = L_i^\mu + \sum_{j=1}^N K_{ij}^\mu L_j^\mu + \sum_{j=1}^N \sum_{k=1}^N K_{ij}^\mu K_{jk}^\mu m_k^\mu. \quad (20)$$

If applied recurrently infinite times, this procedure results in

$$m_i^\mu = L_i^\mu + \sum_{j=1}^N K_{ij}^\mu L_j^\mu + \sum_{j=1}^N \sum_{k=1}^N K_{ij}^\mu K_{jk}^\mu L_k^\mu + \dots \quad (21)$$

which, by exchanging mute variables, can be re-written as

$$m_i^\mu = L_i^\mu + \sum_{j=1}^N L_j^\mu \left\{ K_{ij}^\mu + \sum_{k=1}^N K_{ik}^\mu K_{kj}^\mu + \sum_{k,l=1}^N K_{ik}^\mu K_{kl}^\mu K_{lj}^\mu + \dots \right\}. \quad (22)$$

Eq. 22 can be decomposed into the contribution of the activity of  $G_i$  on one side and that of the rest of the neurons on the other, which will correspond to the first and the second term on the RHS of Eq. 13. To re-obtain this equation we multiply by  $\xi_i^\mu$  and sum over  $\mu$ , using the definition of  $L_i^\mu$  from Eqs. 18,

$$\begin{aligned} \sum_{\mu \neq 1} m_i^\mu \xi_i^\mu &= G_i g_i \sum_{\mu \neq 1} \frac{\xi_i^\mu (1 - a_i)}{Cd_\mu} \left( c_{ii} + \sum_{j=1}^N c_{ji} \left\{ K_{ij}^\mu + \sum_{k=1}^N K_{ik}^\mu K_{kj}^\mu + \dots \right\} \right) + \\ &+ \sum_{l \neq i} G_l g_l \sum_{\mu \neq 1} \frac{\xi_i^\mu (\xi_l^\mu - a_l)}{Cd_\mu} \left( c_{il} + \sum_{j=1}^N c_{jl} \left\{ K_{ij}^\mu + \sum_{k=1}^N K_{ik}^\mu K_{kj}^\mu + \dots \right\} \right). \end{aligned} \quad (23)$$

Let us first treat the first term of Eq. 23, corresponding to  $\gamma_i \sigma_i$  in Eq. 13. Taking into account that  $c_{ii} = 0$  (no self-excitation), only the contribution including the curly brackets survives. As shown in [Roudi and Treves, 2004], each term inside the curly brackets, containing the product of multiple  $K$ 's, is different only to a vanishing order from the product of independent averages, each one corresponding to the sum of  $K_{ab}$  over all pre-synaptic neurons  $b$ . In this way,

$$G_i g_i (1 - a_i) \sum_{\mu \neq 1} \frac{\xi_i^\mu}{C d_\mu} \sum_{j, l_1 \dots l_n = 1}^N c_{ji} K_{il_1}^\mu \left[ \prod_{o=1}^{n-2} K_{l_o l_{o+1}}^\mu \right] K_{l_n j}^\mu \simeq \alpha G_i g_i a_i (1 - a_i) \frac{C}{N} \left\langle \frac{1}{d_\mu^{n+1}} \right\rangle_\mu (a\Omega)^n, \quad (24)$$

where we have introduced  $\alpha \equiv p/C$ , or the memory load normalized by the number of connections per neuron. The  $\langle \dots \rangle_\mu$  brackets symbolize an average over the index  $\mu$  and  $\Omega$  is a variable of order 1 defined by

$$\Omega \equiv \frac{1}{aN} \sum_{j=1}^N a_j (1 - a_j) G'_j g_j. \quad (25)$$

Adding up all the terms with different powers of  $\Omega$  in Eq. 24 results in

$$\gamma_i \sigma_i = \alpha a_i (1 - a_i) g_i \frac{C}{N} \left\langle \frac{\Omega}{d_\mu (d_\mu/a - \Omega)} \right\rangle_\mu G_i. \quad (26)$$

Since  $\Omega$  does not depend on  $\mu$ , if  $d_\mu = a$  for all  $\mu$  the average results simply in the classical  $\Omega/(1 - \Omega)$  factor.

As postulated in the ansatz, the second term in Eq. 23 is a sum of many independent contributions and can thus be thought of as a gaussian noise. Its mean is zero by virtue of the factor  $(\xi_i^\mu - a_i)$ , uncorrelated with both  $\xi_i^\mu$  (by hypothesis) and  $d_\mu$  (negligible correlation). Its variance is given by

$$\langle \langle \rho_i^2 \rangle \rangle = \left\langle \left\langle \sum_{l \neq i} G_l^2 g_l^2 \sum_{\mu \neq 1} \frac{\xi_i^\mu (\xi_l^\mu - a_l)^2}{C^2 d_\mu^2} \left( c_{il} + \sum_{j=1}^N c_{jl} \left\{ K_{ij}^\mu + \sum_{k=1}^N K_{ik}^\mu K_{kj}^\mu + \dots \right\} \right)^2 \right\rangle \right\rangle \quad (27)$$

which corresponds to the first and only surviving term of Eq. 7, the other three terms vanishing for identical reasons. Distributing the square in the big parenthesis and repeating the steps of Eq. 24 this results in

$$\begin{aligned} \langle \langle \rho_i^2 \rangle \rangle &= \alpha a_i \left\{ \left\langle \frac{1}{d_\mu^2} \right\rangle_\mu + 2 \frac{C}{N} \left\langle \frac{\Omega}{d_\mu^2 (d_\mu/a - \Omega)} \right\rangle_\mu + \frac{C}{N} \left\langle \frac{\Omega^2}{d_\mu^2 (d_\mu/a - \Omega)^2} \right\rangle_\mu \right\} \times \\ &\times \sum_{\mu \neq 1} \frac{1}{C} \sum_{l \neq i} (\xi_l^\mu - a_l)^2 g_l^2 \frac{C}{N} G_l^2. \end{aligned} \quad (28)$$

If we define

$$q \equiv \{ \dots \} \frac{1}{N} \sum_{l=1}^N G_l^2 a_l (1 - a_l) g_l^2 \quad (29)$$

including the whole content of the curly brackets from the previous equation, then the variance of the gaussian noise is simply  $\alpha a_i q$ , and the second term of Eq. 13 becomes

$$\rho_i z_i = \sqrt{\alpha a_i q} z_i \quad (30)$$

with  $z_i$ , as before, an independent normally-distributed random variable with unitary variance. The initial hypothesis of Eq. 13 is, thus, self consistent.

Taking into account these two contributions, the mean field experienced by a neuron  $i$  when retrieving pattern 1 is

$$h_i = \xi_i^1 m + \alpha a_i (1 - a_i) G_i g_i \frac{C}{N} \left\langle \frac{\Omega}{d_\mu (d_\mu/a - \Omega)} \right\rangle_\mu + \sqrt{\alpha q a_i} z_i, \quad (31)$$

where we have used  $m_i^1 \simeq m$  and

$$m \equiv \frac{1}{N d_1} \sum_{j=1}^N (\xi_j^1 - a_j) g_j \sigma_j \quad (32)$$

is a variable measuring the weighted overlap between the state of the network and the pattern 1, which together with  $q$  (Eq. 29) and  $\Omega$  (Eq. 25) form the group of macroscopic variables describing the possible stable states of the system. While  $m$  is a variable related to the signal that pushes the activity toward the attractor,  $q$  and  $\Omega$  are noise variables. Diluted connectivity is enough to make the contribution of  $\Omega$  negligible (in which case the diluted equations [Kropff and Treves, 2007] are re-obtained), while  $q$  gives a relevant contribution as long as the memory load is significantly different from zero,  $\alpha = p/C > 0$ .

To simplify the analysis we adopt the zero temperature limit ( $\beta \rightarrow \infty$ ), which turns the sigmoidal function of Eq. 2 into a step function. To obtain the mean activation value of neuron  $i$ , the field  $h_i$  defined by Eq. 31 must be inserted into Eq. 2 and the equation in the variable  $\sigma_i$  solved. This equation is

$$\sigma_i = \Theta \left[ \xi_i^1 m + \alpha a_i (1 - a_i) \sigma_i g_i \frac{C}{N} \left\langle \frac{\Omega}{d_\mu (d_\mu/a - \Omega)} \right\rangle_\mu + \sqrt{\alpha q a_i} z_i - U \right], \quad (33)$$

where  $\Theta[x]$  is the Heaviside function yielding 1 if  $x > 0$  and 0 otherwise. When  $z_i$  has a large enough modulus, its sign determines one of the possible solutions,  $\sigma_i = 1$  or  $\sigma_i = 0$ . However, for a restricted range of values,  $z_- \leq z_i \leq z_+$ , both solutions are possible. Using the definition of  $\gamma_i$  in Eq. 26 to simplify notation, we can write  $z_+ = (U - \xi_i^1 m)/\sqrt{\alpha q a_i}$  and  $z_- = (U - \xi_i^1 m - \gamma_i)/\sqrt{\alpha q a_i}$ . A sort of Maxwell rule must be applied to choose between the two possible solutions [Shiino and Fukai, 1993], by virtue of which the point of transition between the  $\sigma_i = 0$  and the  $\sigma_i = 1$  solutions is the average between the two extremes

$$y_\xi \equiv \frac{z_+ + z_-}{2} = \frac{U - \xi_i^1 m - \gamma_i/2}{\sqrt{\alpha q a_i}}. \quad (34)$$

Inserting Eq. 33 into Eq. 32 yields

$$m = \frac{1}{N d_1} \sum_{j=1}^N (\xi_j^1 - a_j) g_j \int_{-\infty}^{\infty} D z \Theta[z - y_\xi], \quad (35)$$

where we have introduced the average over the independent normal distribution  $Dz$  for  $z_j$ . This expression can be integrated resulting in

$$m = \frac{1}{N d_1} \sum_{j=1}^N (\xi_j^1 - a_j) g_j \phi[y_\xi], \quad (36)$$

where we define

$$\phi(y_\xi) \equiv \frac{1}{2} \left\{ 1 + \operatorname{erf} \left[ \frac{y_\xi}{\sqrt{2}} \right] \right\}. \quad (37)$$

Following the same procedure, Eq. 29 can be rewritten as

$$\begin{aligned} q = & \left\{ \left\langle \frac{1}{d_\mu^2} \right\rangle_\mu + 2 \frac{C}{N} \left\langle \frac{\Omega}{d_\mu^2 (d_\mu/a - \Omega)} \right\rangle_\mu + \frac{C}{N} \left\langle \frac{\Omega^2}{d_\mu^2 (d_\mu/a - \Omega)^2} \right\rangle_\mu \right\} \times \\ & \times \frac{1}{N} \sum_{j=1}^N a_j (1 - a_j) g_j^2 \phi(y_\xi). \end{aligned} \quad (38)$$

Before repeating these steps for the variable  $\Omega$  we note that

$$\int Dz G'_j = \frac{1}{\sqrt{\alpha q a_j}} \int Dz \frac{\partial \sigma_j}{\partial z} = \frac{1}{\sqrt{\alpha q a_j}} \int Dz z \sigma_j, \quad (39)$$

where we have applied integration by parts. Eq. 24 results then in

$$\Omega = \frac{1}{Na} \sum_{j=1}^N \frac{a_j(1-a_j)g_j}{\sqrt{2\pi\alpha q a_j}} \exp \left\{ -\frac{y_\xi^2}{2} \right\}. \quad (40)$$

Eqs. 36, 38 and 40 define the stable states of the network. Retrieval is successful if the stable value of  $m$  is close to 1. In Figure 2 we show the performance of a fully connected network storing the feature norms of McRae and colleagues [McRae et al., 2005] in three situations: theoretical prediction for a diluted network as in [Kropff and Treves, 2007], theoretical prediction for a fully connected network calculated from Eqs. 36-40 and the actual simulations of the network. The figure shows that the fully connected theory better approximates the simulations, performed with random subgroups of patterns of varying size  $p$  and full connectivity for each neuron,  $C = N$ , equal to the total number of features involved in the representation of the subgroup of concepts.

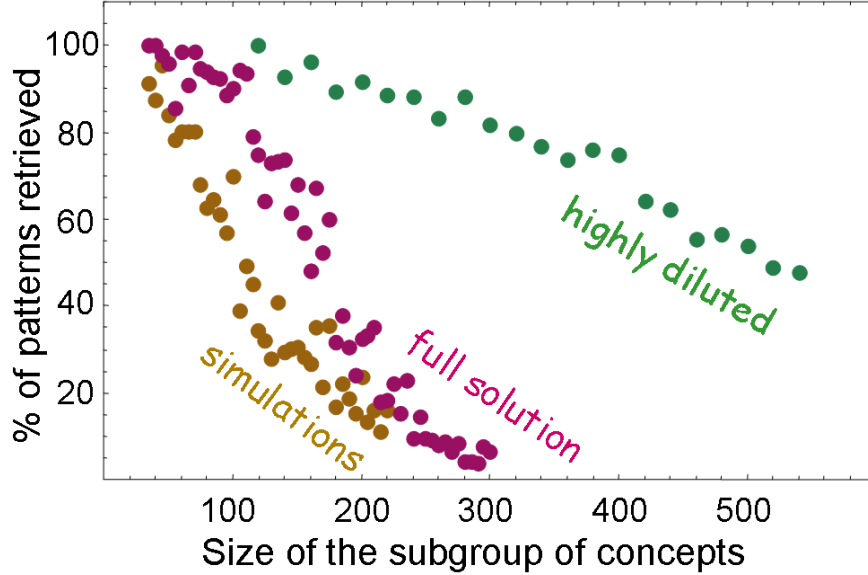


Figure 2: Simulations and numerical solutions of the equations of a network storing random subgroups of patterns taken from the feature norms of McRae and colleagues. The performance of the network depends strongly on the size of the subgroup. Though this is observed in the highly diluted approximation, the decay in performance is not enough to explain the data. It is the full solution with  $g(x) = 1$  that results in a good fit of the simulations. In each simulation, the number of neurons equals the number of features describing some of the stored concepts, and there is full connectivity between neurons,  $C = N$ .

Finally, we can rewrite Eqs. 36-40 in a continuous way by introducing two types of popularity distribution across neurons:

$$F(x) = P(a_i = x) \quad (41)$$

as the global distribution, and

$$f(x) = P(a_i = x | \xi_i^1 = 1) \quad (42)$$

as the distribution related to the pattern that is being retrieved.

The equations describing the stable values of the variables become



$$\begin{aligned}
m &= \int_0^1 f(x)g(x)(1-x)\phi(y_1) - \frac{1}{d_1} \int_0^1 [F(x) - d_1 f(x)] g(x)x\phi(y_0) \\
q &= \left\{ \left\langle \frac{1}{d_\mu^2} \right\rangle_\mu + 2\frac{C}{N} \left\langle \frac{\Omega}{d_\mu^2(d_\mu/a - \Omega)} \right\rangle_\mu + \frac{C}{N} \left\langle \frac{\Omega^2}{d_\mu^2(d_\mu/a - \Omega)^2} \right\rangle_\mu \right\} \times \\
&\quad \times \left\{ d_1 \int_0^1 f(x)g^2(x)x(1-x)\phi(y_1) + \int_0^1 [F(x) - d_1 f(x)] g^2(x)x(1-x)\phi(y_0) \right\} \\
\Omega &= \frac{d_1}{a} \int_0^1 f(x)g(x) \frac{x(1-x)}{\sqrt{2\pi\alpha qx}} \exp(-y_1^2/2) + \frac{1}{a} \int_0^1 [F(x) - d_1 f(x)] g(x) \frac{x(1-x)}{\sqrt{2\pi\alpha qx}} \exp(-y_0^2/2), \quad (43)
\end{aligned}$$

where, adapted from Eq. 34,

$$y_\xi = \frac{1}{\sqrt{\alpha qx}} \left( U - \xi m - \alpha x(1-x)g(x) \frac{C}{2N} \left\langle \frac{\Omega}{d_\mu(d_\mu/a - \Omega)} \right\rangle_\mu \right). \quad (44)$$

## 4 The storage of feature norms

In [Kropff and Treves, 2007] we have shown that the *robustness* of a memory in a highly diluted network is inversely related to the *information* it carries. More specifically, a stored memory needs a minimum number of connections per neuron  $C_{min}$  that is proportional to

$$I_f \equiv \int_0^1 f(x)x(1-x)dx. \quad (45)$$

In this way, if connections are randomly damaged in a network, the most informative memories are selectively lost.

The distribution  $F(x)$  affects the retrievability of all memories. As we have shown in the same paper, it is typically a function with a maximum near  $x = 0$ . The relevant characteristic of  $F(x)$  is its tail for large  $x$ . If  $F(x)$  decays fast enough, the minimal connectivity scales like

$$C_{min} \propto p I_f \log \left[ \frac{I_F}{a I_f} \right], \quad (46)$$

where  $I_F$  corresponds to the same pseudo-information function as in Eq. 45, but using the distribution  $F(x)$ . If  $F(x)$  decays exponentially ( $F(x) \sim \exp(-x/a)$ ), the scaling of the minimal connectivity is the same, with only a different logarithmic correction,

$$C_{min} \propto p I_f \log^2 \left[ \frac{I_F}{a I_f} \right]. \quad (47)$$

The big difference appears when  $F(x)$  has a tail that decays as slow as a power law ( $F(x) \sim x^{-\gamma}$ ). The minimal connectivity is then much larger

$$C_{min} \propto \frac{p I_f}{a} \log \left[ \frac{a^{\gamma-2}}{I_f} \right] \quad (48)$$

since the sparseness, measuring the global activity of the network, is in cortical networks  $a \ll 1$ . Unfortunately, as can be seen in Figure 3, the distribution of popularity  $F(x)$  for the feature norms of McRae and colleagues is of this last type. This is the reason why, as shown in Figure 2, the performance of the network is very poor in storing and retrieving patterns taken from this dataset. In a fully connected network as the one shown in the figure, a stored pattern can be retrieved as long as its minimal connectivity  $C_{min} \leq N$ , the number of connections per neuron. Along the  $x$  axis of the Figure, representing the number of patterns from the norms stored in the network, the average of  $I_f$  is rather constant,  $p$  and  $N$  increase proportionally and  $a$  decreases, eventually taking  $C_{min}$  over the full connectivity limit.

In the following subsections, we analyze different ways to increase this poor storage capacity and effectively store and retrieve the feature norms in an autoassociative memory.

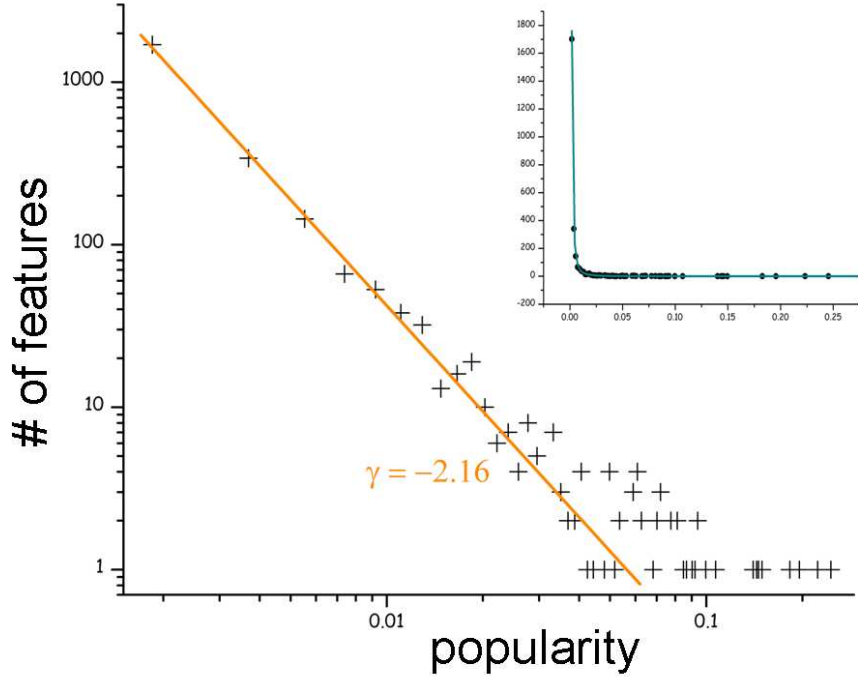


Figure 3: The popularity distribution  $F(x)$  of the feature norms is a power law, with  $\gamma \simeq 2.16$ . Note that both axes are logarithmic. In the inset, the same plot appears with linear axes, including the corresponding fit.

#### 4.1 Adding uninformative neurons

As discussed in [Kropff and Treves, 2007], a way to increase the storage capacity of the network in general terms is to push the distribution  $F(x)$  toward the smaller values of  $x$ . One possibility is to add neurons with low information value (i.e. with low popularity) so as to make  $I_f$  smaller in average without affecting the sparseness  $a$  too much. In Figure 4a we show that the full set of patterns from the feature norms can be stored and retrieved if 5 new neurons per pattern are added, active in that particular pattern and in no other one.

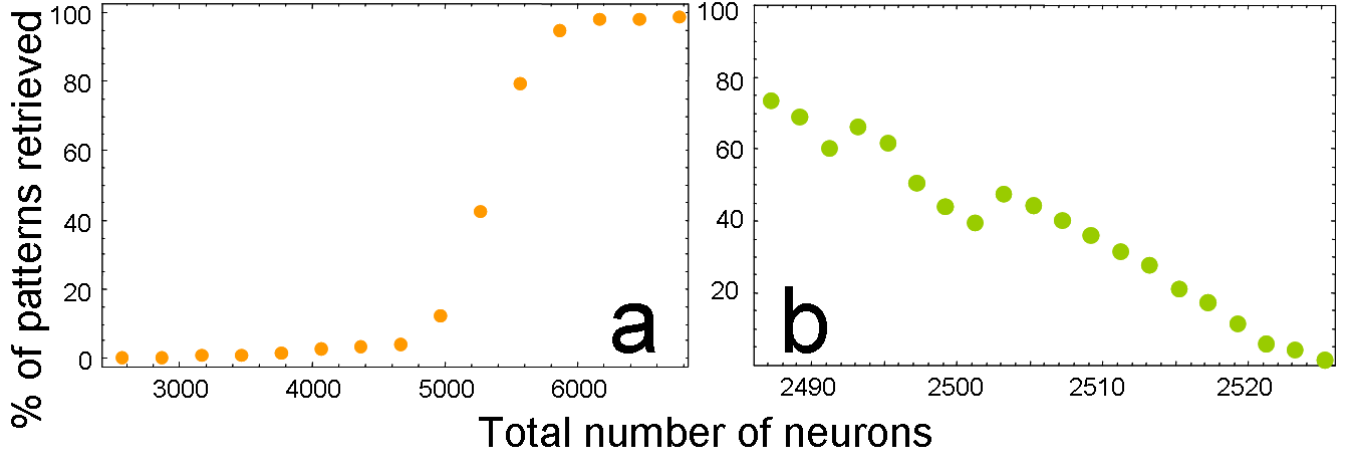
#### 4.2 Removing informative neurons

A similar effect on the distribution  $F(x)$  can be obtained by eliminating selectively the most informative neurons. In Figure 4b we show that if the full set of patterns is stored a retrieval performance of  $\sim 80\%$  is achieved if the 40 more informative features are eliminated. We estimate that 100% performance should be achieved if around 60 neurons were selectively eliminated.

It is not common in the neural literature to find a poor performance that is improved by damaging the network. This must be interpreted in the following way. The connectivity of the network is not enough to sustain the retrieval of the stored patterns, too informative to be stable states of the system. By throwing away information, the system can be brought back to work. However, a price is being payed: the representations are impoverished since they no longer contain the most informative features.

#### 4.3 Popularity-modulated weights

A final way to push the distribution  $F(x)$  toward low values of  $x$  can be figured from Eqs. 43. Indeed,  $g(x)$  can be thought of as a modulator of the distributions  $F(x)$  and  $f(x)$ . Inspired in [Kropff and Treves, 2007], if  $g(x)$  decays exponentially or faster, the storage capacity of a set of patterns with any decaying  $F(x)$  distribution should be brought back to a  $C_{min} \propto pI_f$  dependence, without the  $a^{-1} \gg 1$  factor.



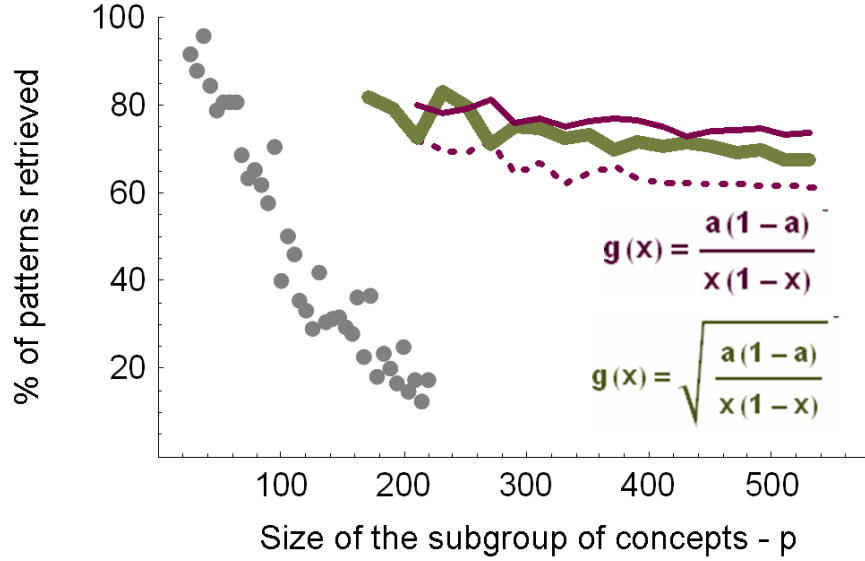


Figure 5: Simulations (dashed line) and theoretical predictions (solid line) of a network storing subgroups of patterns of varying size taken from McRae and colleagues feature norms with a popularity-modulated hebbian learning rule. The thin violet lines use a value of  $g(x)$  inversely proportional to  $x(1-x)$ , normalized so as to maintain the average field of order 1. The thick green line corresponds to a  $g(x)$  inversely proportional to  $\sqrt{x(1-x)}$ . Following our predictions, the exact form of  $g(x)$  does not affect the general performance, which is substantially improved with respect to the simulations with  $g(x) = 1$ , copied from Figure 3 in grey dots.

solution between the two extremes. Whether or not it is a cortical strategy applied to deal with correlated representations is a question for which we have yet no experimental evidence.

## References

- [Amit, 1989] Amit, D. J. (1989). *Modelling Brain Function: the World of Attractor Neural Networks*. Cambridge University Press.
- [Blumenfeld et al., 2006] Blumenfeld, B., Preminger, S., Sagi, D., and Tsodyks, M. (2006). Dynamics of memory representations in networks with novelty-facilitated synaptic plasticity. *Neuron*, 52(2):383–394.
- [Buhmann et al., 1989] Buhmann, J., Divko, R., and Schulten, K. (1989). Associative memory with high information content. *Phys Rev A*, 39:2689–2692.
- [Cree et al., 2006] Cree, G. S., McNorgan, C., and McRae, K. (2006). Distinctive features hold a privileged status in the computation of word meaning : Implications for theories of semantic memory. *Journal of Experimental Psychology*, 32:643–658.
- [Cree et al., 1999] Cree, G. S., McRae, K., and McNorgan, C. (1999). An attractor model of lexical conceptual processing: Simulating semantic priming. *Cognitive Science*, 23(3):371–414.
- [Diederich and Oppen, 1987] Diederich, S. and Oppen, M. (1987). Learning of correlated patterns in spin-glass networks by local learning rules. *Phys. Rev. Lett.*, 58(9):949–952.
- [Gardner et al., 1989] Gardner, E. J., Stroud, N., and Wallace, D. J. (1989). Training with noise and the storage of correlated patterns in a neural network model. *J. Phys. A: Math. Gen.*, 22:2019–2030.

- [Garrard et al., 2001] Garrard, P., Ralph, M. A. L., Hodges, J. R., and Patterson, K. (2001). Prototypicality, distinctiveness, and intercorrelation: Analyses of the semantic attributes of living and nonliving concepts. *Cognitive Neuropsychology*, 18:125 – 174.
- [Gutfreund, 1988] Gutfreund, H. (1988). Neural networks with hierarchically correlated patterns. *Phys. Rev. A*, 37(2):570–577.
- [Hebb, 1949] Hebb, D. (1949). *The organization of behavior*. Wiley: New York.
- [Hopfield, 1982] Hopfield, J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. USA*, 79:2554–2558.
- [Kropff and Treves, 2007] Kropff, E. and Treves, A. (2007). Uninformative memories will prevail: the storage of correlated representations and its consequences. submitted.
- [McRae et al., 2005] McRae, K., Cree, G. S., Seidenberg, M. S., and McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavioral Research Methods, Instruments, and Computers*, 37:547–559.
- [McRae et al., 1997] McRae, K., de Sa, V., and Seidenberg, M. (1997). On the nature and scope of featural representations of word meaning. *Journal of Experimental Psychology: General*, 126(2):99–130.
- [Monasson, 1992] Monasson, R. (1992). Properties of neural networks storing spatially correlated patterns. *J. Phys. A: Math. Gen.*, 25:3701–3720.
- [Parga and Virasoro, 1986] Parga, N. and Virasoro, M. A. (1986). The ultrametric organization of memories in a neural network. *J. Physique*, 47(11):1857–1864.
- [Roudi and Treves, 2004] Roudi, Y. and Treves, A. (2004). An associative network with spatially organized connectivity. *Journal of Statistical Mechanics: Theory and Experiment*, 2004(07):P07010.
- [Shiino and Fukai, 1992] Shiino, M. and Fukai, T. (1992). Self-consistent signal-to-noise analysis and its application to analogue neural network with asymmetric connections. *J. Phys. A*, 25:L375.
- [Shiino and Fukai, 1993] Shiino, M. and Fukai, T. (1993). Self-consistent signal-to-noise analysis of the statistical behavior of analog neural networks and enhancement of the storage capacity. *Phys. Rev. E*, 48:867.
- [Tsodyks and Feigel’Man, 1988] Tsodyks, M. V. and Feigel’Man, M. V. (1988). The enhanced storage capacity in neural networks with low activity level. *Europhysics Letters*, 6:101–105.
- [Vinson and Vigliocco, 2002] Vinson, D. P. and Vigliocco, G. (2002). A semantic analysis of grammatical class impairments: semantic representations of object nouns, action nouns and action verbs. *Journal of Neurolinguistics*, 15:317–351.
- [Virasoro, 1988] Virasoro, M. A. (1988). The effect of synapses destruction on categorization by neural networks. *Europhys. Lett.*, 7(4):293–298.
- [Wills et al., 2005] Wills, T. J., Lever, C., Cacucci, F., Burgess, N., and O’Keefe, J. (2005). Attractor dynamics in the hippocampal representation of the local environment. *Science*, 308(5723):873–876.